AD _____

Award Number: DAMD17-99-1-9174

TITLE: Computer Aid for the Decision to Biopsy Breast Lesions

PRINCIPAL INVESTIGATOR: Carey E. Floyd, Jr., Ph.D.

CONTRACTING ORGANIZATION: Duke University Medical Center
Durham, North Carolina 27710

REPORT DATE: July 2001

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	July 2001	Annual (1 July 2000 – 30 June 2001)

4. TITLE AND SUBTITLE
Computer Aid for the Decision to Biopsy Breast Lesions

5. FUNDING NUMBERS
DAMD17-99-1-9174

6. AUTHOR(S)
Carey E. Floyd, Jr., Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Duke University Medical Center
Durham, North Carolina 27710

E-MAIL:
cef@deckard.mc.duke.edu

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT ***(Maximum 200 Words)***

The goal of this project is to improve the accuracy of the diagnosis of breast cancer from mammograms. In current practice, an estimated 2-10% of true cancers are not biopsied but are followed instead while between 60% and 90% of breast biopsies are performed on benign lesions. This report documents progress so far which demonstrates that this accuracy could be improved by a Case Based Reasoning approach to predict the outcome of a biopsy from the known biopsy outcomes for similar cases. The current version of the CBR performs with an accuracy of 61% on a retrospective set of consecutive cases for which the clinical diagnostic accuracy was 35%. This potential improvement in accuracy potentially improving the accuracy of the diagnosis of breast cancer from 35%to 61% for a set of 1023 cases. The CBR algorithm has been examined through evaluating and refining different techniques for the four fundamental tasks 1)specify a reference set of cases; 2)define a metric for the distance between cases; 3)define a rule (based on the distance metric) for selecting "similar" cases from the reference set; 4)specify a classification technique for predicting the outcome of biopsy from the known outcomes of the selected similar reference cases.

14. SUBJECT TERMS
Breast Cancer Computer-Aided Diagnosis

15. NUMBER OF PAGES
17

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the U.S. Army.

N/A Where copyrighted material is quoted, permission has been obtained to use such material.

N/A Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

N/A Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products or services of these organizations.

N/A In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and use of Laboratory Animals of the Institute of Laboratory Resources, national Research Council (NIH Publication No. 86-23, Revised 1985).

N/A For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal law 45 CFR 46.

N/A In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institute of Health.

N/A In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA molecules.

N/A In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.

Carey E. [illegible] 30 Oct 02

PI – Signature Date

Table of Contents

Report of the Progress on Grant DAMD17-99-1-9174

For the Period of July 2000 to June 2001

Computer Aid for the Decision to Biopsy Breast Lesions

Hypothesis:

This work will test the hypothesis: "The results of breast biopsy can be accurately predicted from the results of biopsies for previous cases that had similar mammographic abnormalities."

Nomenclature

To clarify a potential source of confusion in this proposal, two terms are defined here: "feature" and "finding". The term "feature" refers to a variable while the term 'finding' refers to the value of a variable. For the categorical descriptions of mammographic abnormalities described below in Table 5, an example would be "the feature 'mass margin' has a finding of 'spiculated'.

Significance for reducing the number of benign biopsies

The lifetime risk of developing breast cancer has increased steadily from 1940, when the first statistics were collected, to the present risk of one woman in eight [1]. Several large studies have demonstrated that screening mammography can decrease the mortality due to breast cancer by 30%[2, 3]. Unfortunately, evaluating mammograms is a complicated task. Multiple radiographic features of each mammographic abnormality must be examined to determine whether further action such as follow-up or biopsy for histologic diagnosis is warranted. Although mammography is a sensitive tool for detecting breast cancer, the positive predictive value (PPV) has historically been low [4-6]. Due to the overlap of the radiographic appearance of benign and malignant breast lesions [6] as well as an overall conservative approach of physicians [7], only about 20% of women who undergo biopsy for mammographically suspicious non-palpable lesions have a malignancy by histologic diagnosis[5]. This relatively low Positive Predictive Value of mammography-induced biopsy is recognized as a significant problem. If the mammography screening recommendations of the American College of Radiology (ACR) and the American Cancer Society (ACS) are fully implemented, nearly all women over the age of 40 will undergo a yearly mammogram. Currently, the biopsy rate is 0.5 - 2.0% of all mammographic exams. Potentially, several million biopsies will be performed each year [8]. With the current accuracy, hundreds of thousands of women who do not have breast cancer would be unnecessarily subjected to the discomfort, expense, potential complications, change in cosmetic appearance, and anxiety that can accompany breast biopsy [5, 9-11]. In addition, the financial burden of these procedures (between $3500 and $5000 for excisional and between $1000 and $1500 for core biopsy) is substantial (around $100,000,000 per year)[5, 8, 9]. This project will develop an accurate computer-based system to provide a second opinion to assist the mammographer with the decision to biopsy.

The interpretation and decision process for a diagnostic mammogram is quite different from the screening mammogram. As a second reader in diagnostic mammography, the system could provide a mammographer with 1)a diagnosis, 2) an estimate of uncertainty for the diagnosis, and 3)sample images from the set that were accepted as similar. The mammographer can use this additional information for the decision to recommend biopsy or follow-up. A significant value to the clinician is that the decision aid potentially contains information derived from more cases than any mammographer could have ever seen and thus provides access to an experience base that would not otherwise be available.

The anticipated clinical impact of this CBR second opinion will be to increase the diagnostic accuracy of mammography for predicting malignancy of breast lesions. This will be achieved by decreasing the number of patients sent to biopsy with benign lesions and by decreasing the variability of diagnosis for mammography.

During this project period, studies have been conducted to demonstrate feasibility of this concept. Presented below are results demonstrating feasibility of CBR as a predictive model for the outcome of breast biopsy.

Key Research Accomplishments

1 Analyzed distribution of findings in the case database
2 Non-parametric ROC evaluation of the classifier performance was performed
3 Feature selection was examined for Hamming distance
4 Hamming distance metric was evaluated

A CBR system was developed to classify cases referred for biopsy. The CBR was evaluated on a set of 500 cases from Duke (described in more detail below) using round-robin sampling. All cases were referred to excisional biopsy and the truth for evaluating the classification of each case was abstracted from the pathology report. Of these 500 diagnostic mammography cases that were that were referred to biopsy, 326 (64%) were benign. While this fraction is higher than the value of 20% typically quoted as a national average, it is consistent with that seen at other teaching hospitals. In the framework of the specific aims of this proposal, the properties of this CBR include:

Table 1 Characteristics of CBR used in feasibility studies

Reference data	500 Retrospective biopsy cases from Duke
Case Encoding	Uniformly scaled rank order (see table 3 below)
Similarity Metric	Hamming Distance
Similarity Selection	Threshold applied to Hamming distance metric

Classification metric	Probability of malignancy

Table 1: Characteristics that define the CBR.

Analyzed distribution of findings in the case database

Here we present some characteristics of the reference database that has been acquired. The database consisted of cases that were evaluated at diagnostic mammography after being called back due to an abnormality observed in a screening examination. All of the cases were non-palpable and were referred to biopsy. Cases were excluded if a previous biopsy or surgery had been performed at the site of the abnormality. Outcomes were established from the pathology report. Each case included 1) the mammographers' description of the abnormality using the BI-RADS™ lexicon, 2) known epidemiological risk factors for breast cancer; and 3) outcomes in the form of benign or malignant status as determined by biopsy. The risk factors are routinely acquired by a short patient interview conducted by mammography technologists at the time of the diagnostic examination. Of the 500 lesions evaluated in the feasibility studies, there were 232 masses alone, 192 microcalcifications alone, and 29 combinations of masses and associated microcalcifications. The remaining 47 lesions included various combinations of architectural distortion, regions of asymmetric breast density, areas of focal asymmetric density, and areas of asymmetric breast tissue. Patient age ranged from 24 to 86 years with a mean value of 55 years. At biopsy, 326 (65%) of the lesions were found to be benign while 174 (35%) were found to be malignant. Currently (as of May 2001), our database contains around 1500 cases that were examined at diagnostic mammography and were referred to biopsy at Duke University Medical Center between 1992 and 2000. While this does not represent all of the consecutive cases, the omissions are believed to be random and these data are considered to represent an unbiased sample of the population of cases to which the decision system would be applied.

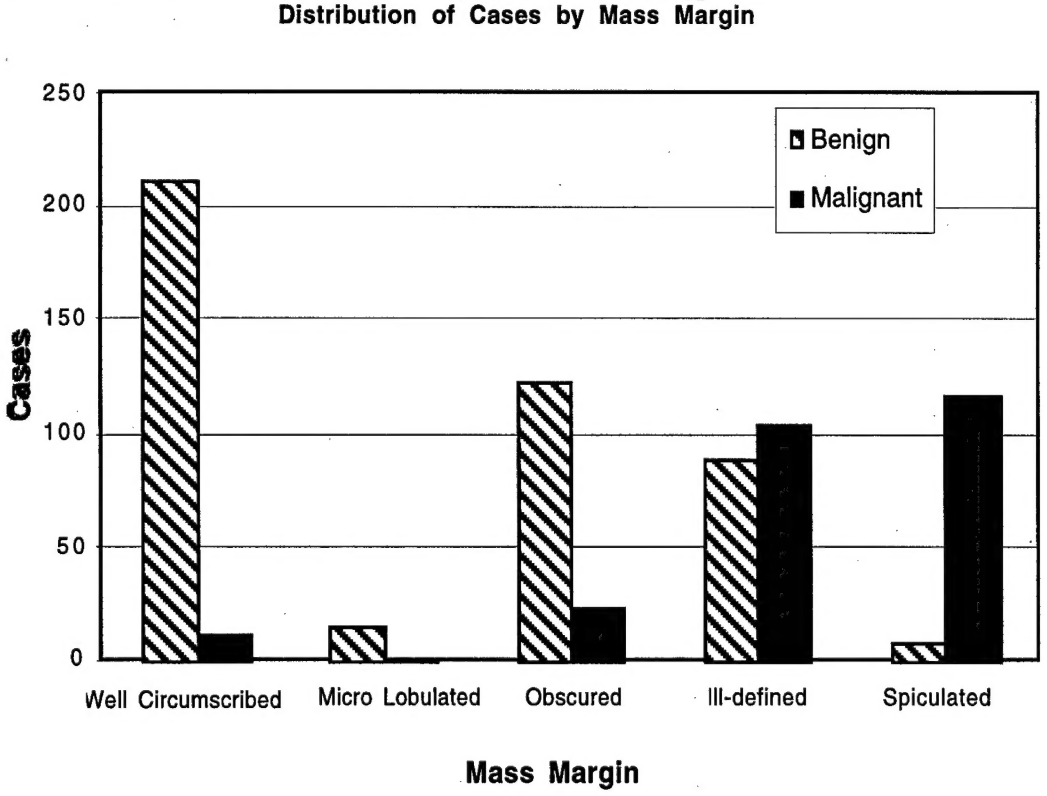

Fig. 4 The distribution of cases by mass margin is shown with malignant cases represented by the dark bars and the benign cases represented by the striped bars.

However, as combinations of features are considered, there is clear evidence that several of the features are not independent and that several of the joint probability distributions are not well determined. From our experience with the parametric fitting described above, we feel that it is important to avoid parametric assumptions where possible in this problem and propose to acquire more cases. When these joint distributions are examined from the data, numerous discontinuities are evident that are believed to be the result of too few cases. This is a concern for a CBR since these distributions are estimated directly from the reference data. As an example, consider the distribution of categorical findings for the mass margin and mass shape features shown in Fig. 4 and Fig. 5. Two observations are evident from Fig. 4. First, masses with Micro Lobulated margins are rarely referred to biopsy. Second, ignoring the Micro Lobulated category, the distributions for benign and malignant cases are monotonicly decreasing/increasing respectively with the findings ordered as shown (which is consistent with the BI-RADS^{TM} specification). From Fig. 4 there seem to be a sufficient number of cases to describe these distributions. In Fig.5 is the distribution of the Mass Shape feature. Here the distributions are not monotonic but the shape is still rather well defined although the relationship between the first three categories for benign masses is uncertain. When the dependence on mass shape is also considered, as shown in Table 2, it is clear that 1)these two features are not independent and 2)the form of the dependence is not well determined with the current number of cases, particularly for the benign masses.

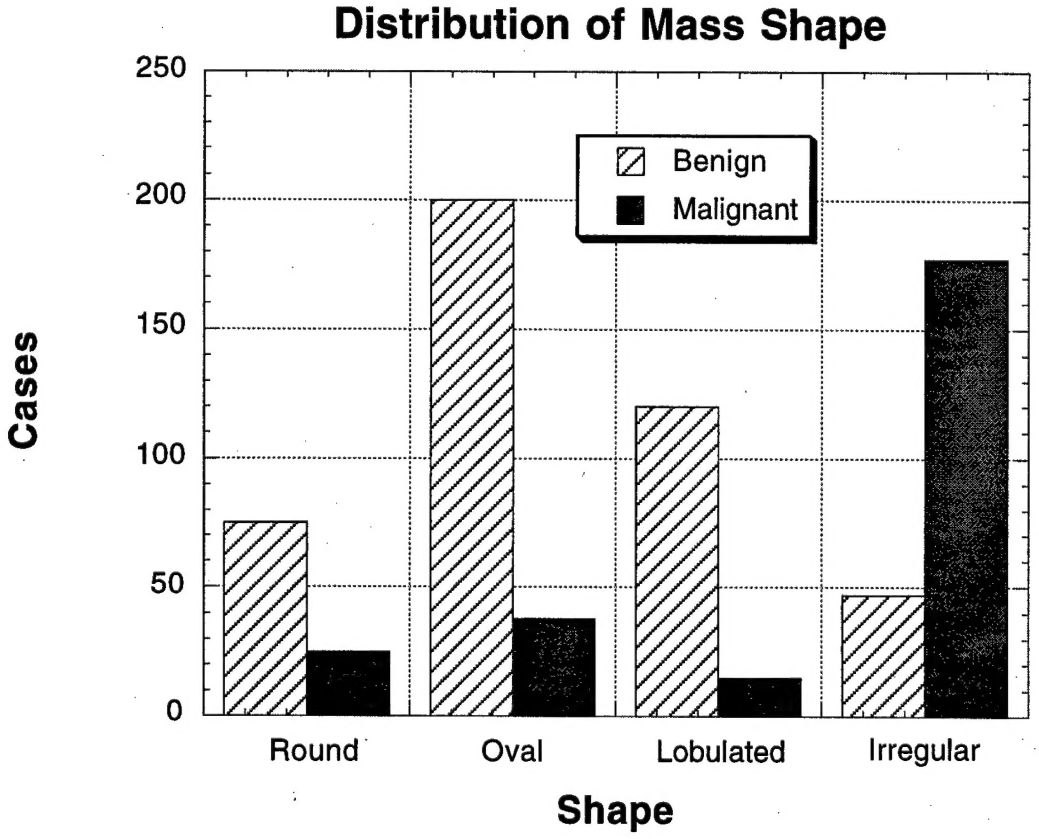

Fig. 5 The distribution of cases by mass shape is shown with malignant cases represented by the dark bars and the benign cases represented by the striped bars.

	Margin	Well-Circumscribed	Micro-Lobulated	ill-defined	Obscured	Spiculated
Shape	round	38	2	22	12	1
Benign	oval	106	5	59	29	1
	lobulated	65	5	34	16	0
	irregular	1	2	7	31	6
	round	2	0	3	17	3
Malignant	oval	7	0	9	19	3
	lobulated	2	1	5	7	0
	irregular	0	0	6	61	110

Table 2 The joint distribution of cases by mass margin and mass shape is shown in this table with malignant cases in the lower pane and benign cases in the upper pane.

Define mathematical representation of a case

We have examined choices for the representation of the cases beginning with the representation used to develop our ANN classifiers. Cases are represented by a vector of features each of which has a number of possible categorical values or findings. BI-RADS™ was developed as a reporting lexicon and not as a direct indicator of probability for disease and while the assignment of numerical values to the categories is not provided, the lexicon does describe a rank order among many of the findings. From this in combination with discussions with several mammographers, a weighting (or value) scale has been developed and used successfully in the previous CBR and ANN analysis. These weights are presented with the findings in Table 5 below. Values were assigned as normalized rank orderings of the categorical values in each finding independently and were intended to rank the possible descriptions in order of their likelihood of malignancy.

Feature selection was examined for Hamming distance

We examined the sensitivity of the CBR to the method used for selecting cases. The selection rule is a combination of a distance metric and a threshold technique. Here several sets of features were examined for computing the Hamming distance and the cutoff threshold was varied. Of interest is the observation that performance increased when the distance increased from 0 (which required an exact match) to 1 (which allowed one of the 6 features to differ). The best performance was found when only three features were required and up to one was allowed to differ. We believe that better performance will be obtained with more than three features but that this will require more cases. This seems likely when considering that with these three features: Mass Margin, Calcification Description and Age, only cases with calcified masses (10% of the cases) could possibly non-null findings for all three features. As a side note, while the best CBR performance is slightly less than the best ANN performance on these cases, the ANN performance is close to chance if only three features are provided.

Hamming distance metric was evaluated

Table 3 Case Based Reasoning: Performance for Hamming Distance						
Number of Features	Feature set	Distance Threshold	ROC Area	Partial ROC Area	Specificity at 100% Sensitivity	Specificity at 98% Sensitivity
6	A	0	0.70	<0.05	<0.01	<0.01
6	A	1	0.79	0.2	<0.01	<0.01
3	B	1	0.83	0.45	0.25	0.41

Table 3. Performance of CBR with Hamming distance as a function of distance threshold and features sets Feature set A: Age, Mass Margin, Mass Shape, Calcification Description, Calcification Distribution, Associated Findings; set B: Age, Mass Margin, Calcification Description.

Table 4 Performance for different thresholds on the probability of malignancy					
Probability Threshold	Sensitivity (%)	Specificity (%)	Positive Predictive Value (%)	Benign Biopsies Avoided	Malignancies Missed
Accept All Cases	100	0	35	0	0
0.10	100	25	42	81	0
0.21	98	41	46	134	10

Table 4. Performance of Case Based Reasoning System for different thresholds applied to the predicted probability of malignancy.

Table 5 - Input Features for breast biopsy cases

BI-RADS™ Lesion Descriptors	**BI-RADS™ Lesion Descriptors**

Input Node	Feature	Finding	Value
1	Calcification Distribution		
		no calcifications	0
		diffuse	0.2
		regional	0.4
		segmental	0.6
		linear	0.8
		clustered	1.0
2	Calcification Number		
		no calcifications	0
		< 5	0.33
		5 to 10	0.66
		> 10	1.0
3	Calcification Description		
		no calcifications	0
	Benign-like findings	milk of calcium-like	0.2
		rim	0.2
		skin	0.2
		vascular	0.2
		spherical	0.2
		suture	0.2
		coarse	0.2
		large rod-like	0.2
		round	0.2
	other	dystrophic	0.4
		punctate	0.6
		indistinct	0.8
		pleomorphic	0.9
		fine branching	1.0
4	Mass Margin		
		no mass	0
		well circumscribed	0.2
		microlobulated	0.4
		obscured	0.6
8	Location		
		___o'clock	
		axillary tail	0
		posterior	0.2
		middle	0.4
		anterior	0.6
		subareolar	0.8
		central	1.0
9	Associated Findings		
		none	0.00
		skin lesion	0.13
		hematoma	0.25
		trabecular thickening	0.38
		nipple retraction	0.50
		skin retraction	0.63
		skin thickening	0.75
		architectural distortion	0.88
		axillary adenopathy	1.00
10	Special Cases		
		none	0
		intramammary lymph node	0.25
		asymmetric breast tissue	0.5
		focal asymmetric density	0.75
		tubular density or solitary dilated duct	1.0

Features Involving Personal and Family History

Input Node	Feature	Finding	Value
11	Age		in years
12	Personal History of breast cancer	none	0
		positive	1

		ill-defined	0.8	13	History of Prior Ipsilateral Benign Biopsy	none	0
		spiculated	1.0			positive	1
				14	Family History of breast cancer	none	0
5	Mass Shape					positive	1
		no mass	0				
		round	0.25	15	Menstrual History	pre-menopausal	0
		oval	0.5			post-menopausal	1
		lobulated	0.75				
		irregular	1.0	16	Estrogen/Progesterone Therapy	none	0
						positive	1
6	Mass Density						
		no mass	0				
		fat-containing	0.25				
		low density	0.5				
		isodense	0.75				
		high density	1.0				
7	Mass Size		mm				

Table 5 shows the case representation that was evaluated. The "value" shown indicates the quantitative values assigned to individual findings in the preliminary data. These were initially assigned by uniformly distributing the rank-ordered findings between 0 and 1 for each feature.

Non-parametric ROC evaluation of the classifier performance was performed

Typically, published ROC curves are smooth since they are obtained though a parametric representation of the data. For a five-category human observer response experiment, this parameterization is necessary and is usually obtained using the software developed by Dr. Charles Metz of the University of Chicago. In the initial experiments, we found that the fitted curves did not accurately follow our data in the regions of high sensitivity which is exactly where we have the most interest in comparing techniques. Outputs of this CBR, the histogram of the negative cases followed a distribution that did not appear to be normal. In particular, the . After consulting with Dr. Metz, we decided that a non-parametric evaluation of the ROC performance would be more appropriate for these data. The source of our difficulty lay in the deviations from the normal distribution that are found in the tails of the probability density functions from the CBR. Interestingly, the ROC area estimates agreed very well, but the shapes were different. For this reason, all ROC curves are presented in a non-parametric form. That is, they are plotted from the data rather than from a fit to the data. With 500 or more continuous valued outputs, the Trapezoid Rule for computing the area gives sufficient accuracy. A convenience of the parametric fitting software is that they provide an estimate of the significance of any differences in performance for paired data. To estimate the significance of a difference computed with non-parametric methods, the mean values and variances (including covariances) for all performance

measures were obtained by bootstrap sampling [49]. For the results presented here, 3000 samples were found to provide asymptotically stable estimates for all performance measurements.

Performance was evaluated by the receiver operating characteristic curve (ROC), the Partial Area Index ($_{0.90}A_z$) computed as the ROC area (scaled by 10) for sensitivities greater than 90%, and the specificity at 98% sensitivity. The ROC curve is shown in Fig. 1 below. Particularly encouraging is the behavior of the curve at high sensitivity, seen more clearly in the plot of $_{0.90}A_z$ in Fig. 2. The sensitivity remains very high as the false positive fraction (FPF) decreases. The sensitivity does not decrease below 98% until the FPF has dropped to 0.6 (specificity of 0.4). At this operating point, 130 of the 326 benign biopsies could be avoided with delayed diagnosis for only two malignant cases.

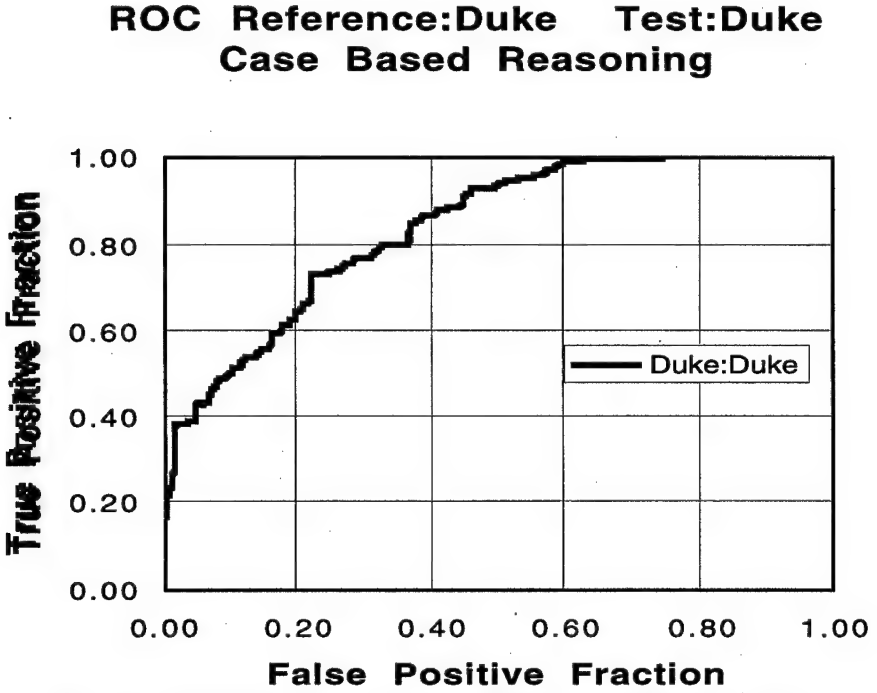

Fig. 1. Full ROC curve for the CBR described in Table 1

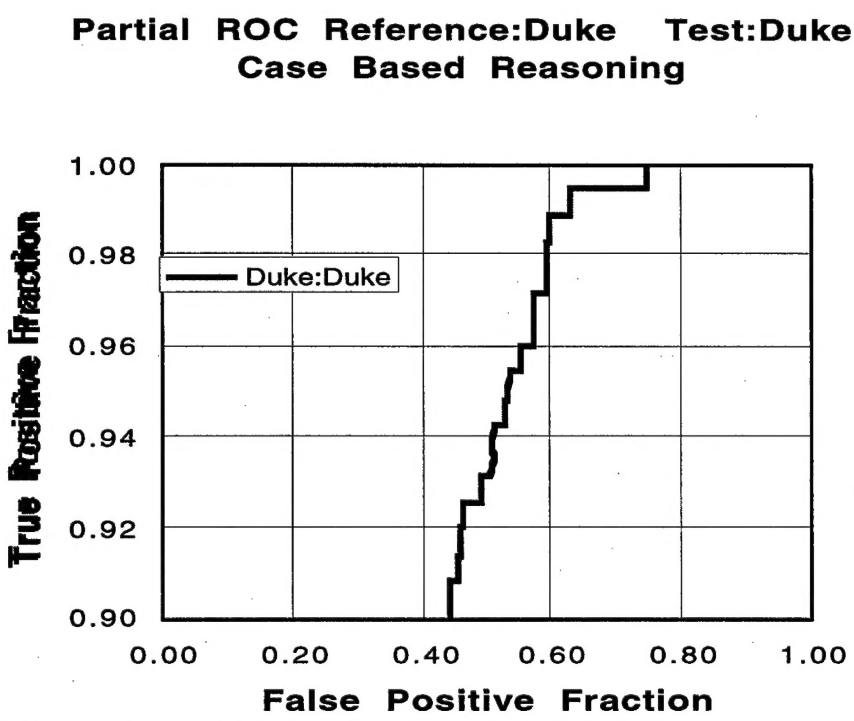

Fig. 2. Partial ROC curve for the CBR described in Table 1. This performance measure is of more clinical relevance than the full ROC for this cancer diagnosis task.

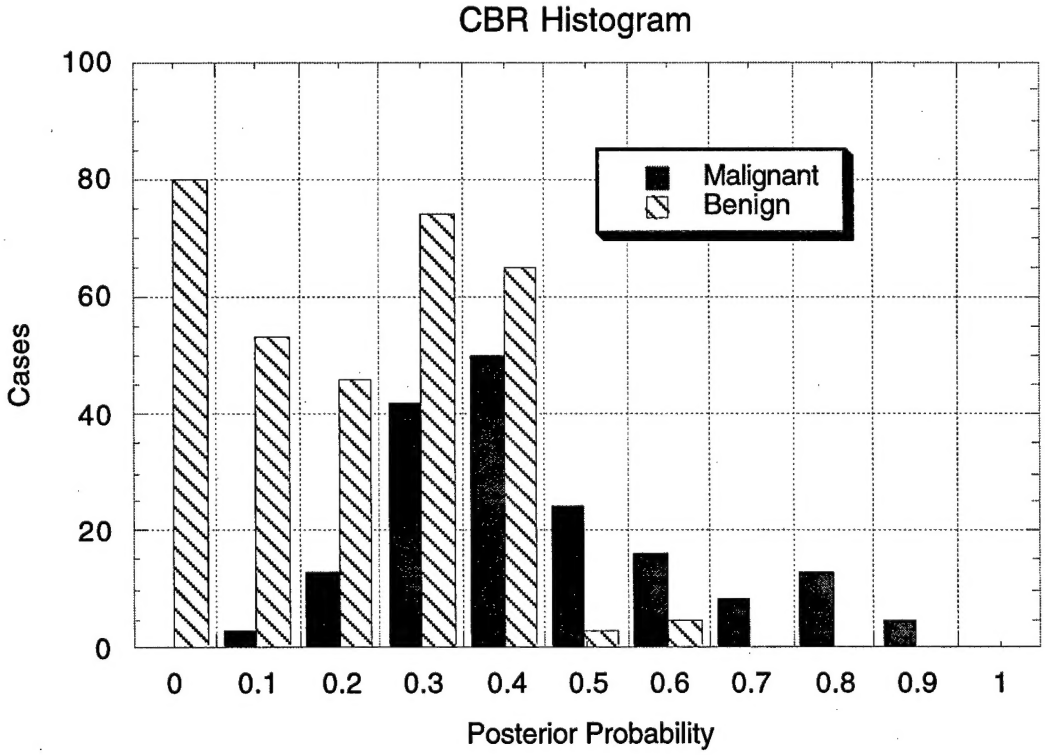

Fig. 3 The histogram of benign and malignant cases for the full range of the CBR output. The benign cases are represented by the striped open bars while the malignant test cases are represented by the gray bars.

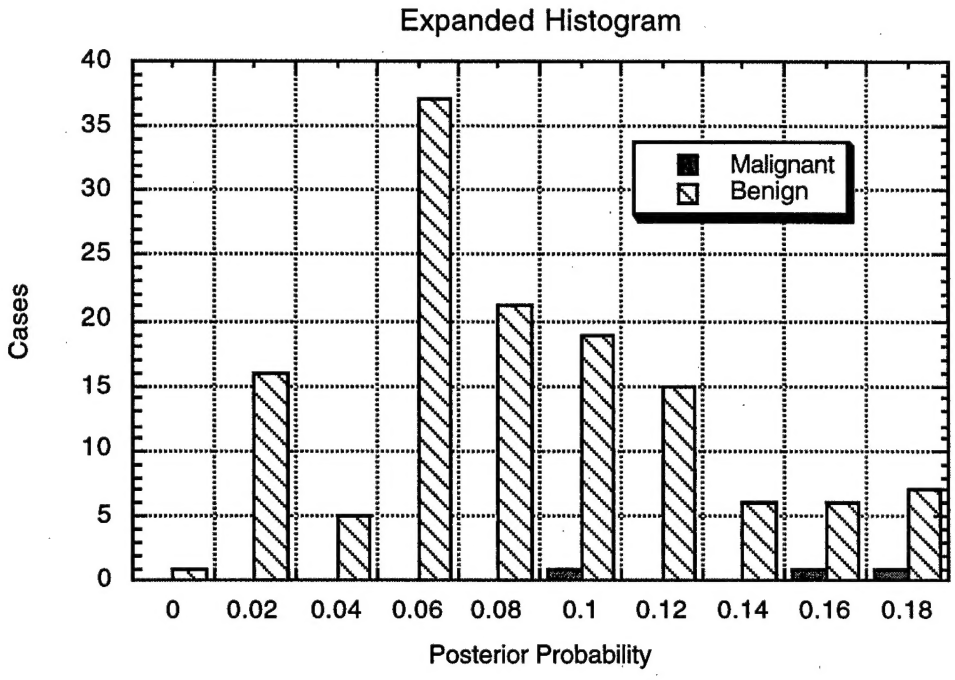

Fig. 4 The histogram of benign and malignant cases for an expanded region of low probability for malignancy.

As seen in Fig. 3, moderate separation of the benign and malignant cases was achieved resulting in an ROC area of 0.83. Since there are some benign cases to the left of all malignant cases, biopsy could be avoided for these without missing any of the malignancies. To further examine this region, the histogram is expanded in Fig. 4 for the region assigned low probability of malignancy.
In this low probability region, there are 133 benign cases and only three malignant cases. The benign cases are represented by the striped open bars while the malignant test cases are represented by the gray bars.

The portion of the ROC curve that is of greatest interest is the region of greatest true-positive fraction (i.e. highest sensitivity) since few radiologists or patients would be willing to miss a larger fraction of breast cancers for the sake of high specificity. The cases populating this region are those that were assigned the lowest probability of malignancy.

It is interesting to note that the cancer shown farthest to the left in Fig. 4 is a 45 year old woman with a small well-circumscribed mass. These characteristics all would indicate a benign mass and the CBR agreed. The critical information that was not included in the model is that this mass was not seen in a previous mammogram. This information will be included in the proposed studies. In addition, it is interesting to note the features of the benign lesions that were assigned a probability lower than any of the malignancies. These are all masses and include 60 with well circumscribed margins, and one mass with a well circumscribed margin and with associated calcifications described as indistinct, one mass with a microlobulated margin, 18 masses with obscured margins, and one mass with an ill-defined margin and with associated calicifications described as coarse.

At sensitivity of 0.98 (relative to all biopsied lesions) the specificity is 0.4. Thus, 40% of the benign biopsies could have been avoided at the cost of delaying diagnosis for 2% of the malignancies. The positive predictive value for these data would be increased from 35% to 46%. These results demonstrate feasibility for developing CBR as a decision aid for breast biopsy using the BI-RADS™ lexicon to index the cases.

Reportable Outcomes

Peer-Reviewed Publications: (Submitted)
A.O. Bilska-Wolak, C.E. Floyd Jr., "Development and evaluation of a case-based reasoning classifier for prediction of breast biopsy outcome with BI-RADS ™ lexicon." Med. Phys.

Conference Proceedings:
A.O. Bilska, C.E. Floyd, Jr, "Investigating different similarity measures for a case-based reasoning classifier to
predict breast cancer," SPIE Vol. 4322, p. 1862-1866, 2001.

Database developed for BIRADS findings of cases referred to biopsy.

Funding applied for from NIH through the R01 mechanism June 2001.

Conclusion

In conclusion, the database was analyzed and the distribution of several features was reported, non-parametric evaluation techniques were explored and found to be more appropriate than parametric techniques, the performance of the CBR classifier was examined under variations of several of the key components of the system. The performance was evaluated for different sets of test data and database data from different institutions. Differences in performance were observed. Performance was evaluated under different sets of input findings and an optimal set was selected. Performance was evaluated under different implementations of the Hamming distance criteria under different cutoff distances. Differences were observed and an optimal cutoff was discovered. These interim results suggest that the current study plan is appropriate and that the CBR approach can be developed into a clinically usable decision tool.

REFERENCES

1 L. Garfinkel, C. C. Boring, and C. W. Heath, "Changing trends: an overview of breast cancer incidence and mortality," Cancer **74**, 222-227 (1994).

2 S. Shapiro, "Screening: assessment of current studies," Cancer **74**, 231-238 (1994).

3 A. L. M. Verbeek, J. H. C. L. Hendriks, R. Holland, M. Mravunac, F. Sturmans, and N. E. Day, "Reduction of breast cancer mortality through mass screening with modern mammography," Lancet **1**, 1222-1224 (1984).

4 D. D. Adler, and M. A. Helvie, "Mammographic biopsy recommendations," Current Opinion in Radiology **4**, 123-129 (1992).

5 D. B. Kopans, "The positive predictive value of mammography," AJR Am J Roentgenol **158**, 521-526 (1992).

6 S. Ciatto, L. Cataliotti, and V. Distante, "Nonpalpable lesions detected with mammography: review of 512 consecutive cases," Radiology **165**, 99-102 (1987).

7 F. M. Hall, "Screening mammography - potential problems on the horizon," NEJM **314**, 53-55 (1986).

8 F. M. Hall, J. M. Storella, D. Z. Silverstone, and G. Wyshak, "Nonpalpable breast lesions: recommendations for biopsy based on suspicion of carcinoma at mammography," Radiology **167**, 353-358 (1988).

9 G. F. Schwartz, D. L. Carter, E. F. Conant, F. H. Gannon, G. C. Finkel, and S. A. Feig, "Mammographically detected breast cancer: nonpalpable is not a synonym for inconsequential," Cancer **73**, 1660-1665 (1994).

10 M. A. Helvie, D. M. Ikeda, and D. D. Adler, "Localization and needle aspiration of breast lesions: complications in 370 cases," AJR Am J Roentgenol **157**, 711-714 (1991).

11 J. M. Dixon, and T. G. John, "Morbidity after breast biopsy for benign disease in a screened population," Lancet **1**, 128 (1992).